MonaLIA 1.0 preliminary study on the coupling learning and reasoning for image recognition to enrich the records of in the Joconde database

Anna Bobasheva

MONALIA 1.0 PROJECT REPORT

Preliminary study on image recognition of the Joconde database in connection with semantic data (JocondeLab)

by Anna Bobasheva

Title (EN) : preliminary study on the coupling learning and reasoning for image recognition to enrich the records of in the Joconde database

Abstract (EN): The MonaLIA 1.0 project is a preliminary study on the coupling of learning methods (Deep Neural Networks) and knowledge-based methods (Semantic Web) for image recognition and the enhancement of descriptive documentary records. The approach is applied and evaluated on the collection and data in the Joconde database in order to identify the possibilities and challenges offered by this coupling in assisting in the creation and maintenance of such an annotated collection.

Title (FR) : étude préliminaire sur le couplage apprentissage-raisonnement pour la reconnaissance d'images et l'enrichissement de notices de la base Joconde

Abstract (FR) : Le projet MonaLIA 1.0 est une étude préliminaire sur le couplage de méthodes d'apprentissage (Réseaux de Neurones Profonds) et de méthodes à base de connaissances (Web Sémantique) pour la reconnaissance d'images et l'enrichissement de notices descriptives documentaires. L'approche est appliquée et évaluée sur la collection et les données de la base Joconde afin d'identifier les possibilités et les verrous offerte par ce couplage dans l'assistance à la création et la maintenance d'une telle collection annotée.

Team at Inria :

- Fabien Gandon (Inria, Université Côte d'Azur, CNRS, I3S), project leader for Inria
- Frédéric Precioso (I3S, Université Côte d'Azur, CNRS)
- Anna Bobasheva (Inria, Université Côte d'Azur, CNRS, I3S), author of this report

Team at French Ministry of Culture (MiC) :

- Laurent Manœuvre (MIC / BDNC), project leader for MIC
- Bertrand Sajus (MIC / DIN)

OBJECTIVE

The goal of this project is to exploit the cross-fertilization of recent advances in image recognition and semantic indexing on annotated image databases in order to improve the accuracy and the details of the annotation. The idea would be, at first, to assess the potential of machine learning (including deep learning) and the semantic annotations on the Joconde database (350 000 illustrated artwork records from French museums). Joconde also contains metadata based on a thesaurus. In a previous project (JocondeLab) these metadata was formalized in semantic Web formalism and is linking the iconographic Garnier thesaurus and DBpedia to the data of the Joconde database.

This project enters into the current emerging trend in deep learning: Deep Reasoning (http://www.college-de-france.fr/site/yann-lecun/seminar-2016-04-08-12h00.htm). This topic focuses on combining the strength of two very powerful but different approaches for extracting information and knowledge from data, deep learning from unstructured data and reasoning from the structured data.

Objectives of the project:

This study would assess the interest and feasibility of several experimental objectives:

- Enhance the records with some keywords from the Garnier thesaurus and with some Inferences on the vocabularies used.
- Use learning based on semantic queries on JocondeLab database to (1) enhance and complete the database (2) apply the same process to other corpuses than Joconde.
- Use pattern recognition to generate a ranking by iconographic relevance, in the lists of results from Joconde; currently with keyword indexing, we only know if an image contains a topic or not, but we cannot assess the importance of the topic compared to the global content of the image: it is unclear if it is the main theme, or just a detail. Accordingly, the resulting lists are polluted by images whose content reported by a keyword is actually anecdotal.

APPROACH

IMAGE SUBSET FOR CATEGORY EXTRACTION

Query (SPARQL) the metadata to extract the subset of images that belong to a certain category or the subcategories below it to train the Neural Network classifier

- Joconde dataset is represented by the RDF files containing artwork records metadata and thesauri (Joconde KB) as well as the collection of the JPEG image files. Metadata provides the association between the artwork records and the image file path in the collection. Each artwork record has a unique ID *noticeRef* that can be used to link the image and the metadata.

- To label the images for the classification I used two fields with hierarchical thesauri: represented subject (REPR thesaurus) and domain (DOMN thesaurus). The former is of the obvious interest and the latter is chosen as an experimental for its relatively simple thesaurus and image quantity sufficiency.

	All records	**Records with images**[1]	**Records with existing images**	
Count	483297	298597	285144	
	Unique terms	**Unique terms**		**Total unique terms count**
REPR terms	29185	22552		37279
DOMN terms	131	129		150

- Query the Joconde KB to extract the records for a specified top category (concept, term) and all its subcategories capturing the hierarchy based on the thesauri.
- Filter the records with available images

PRE-PROCESS THE QUERY RESULTS

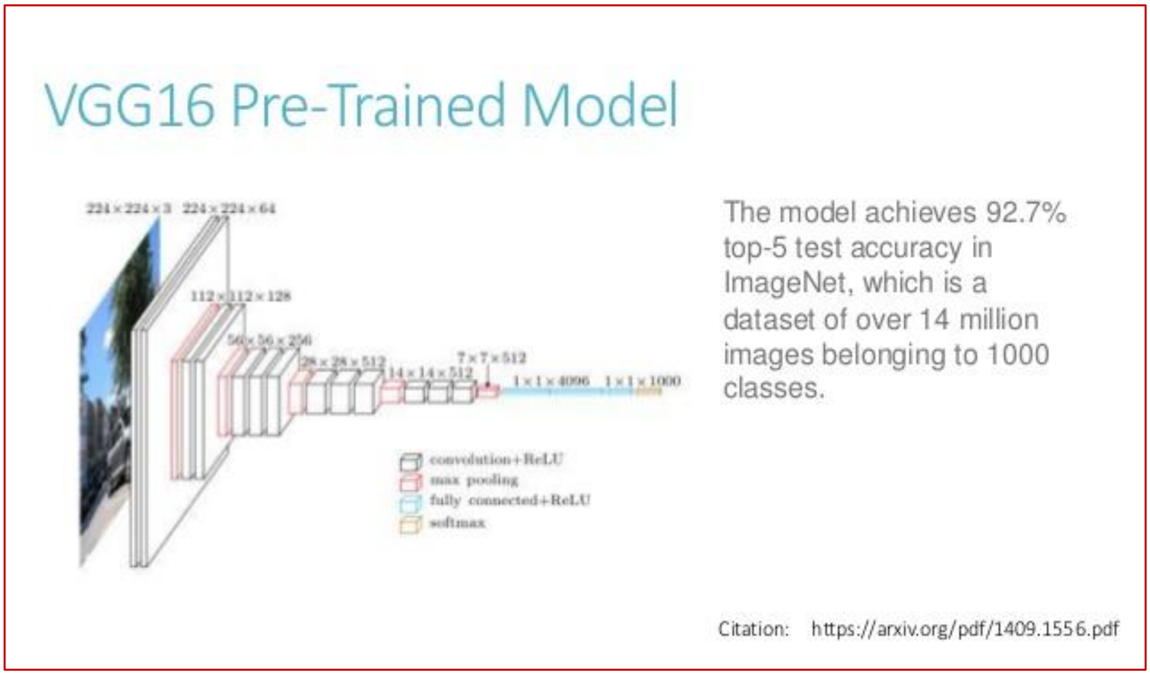
The images come in a proprietary file structure with artwork records containing the reference to the relative path in this structure. The classification framework however requires a different file structure so the image files have to be selected, rearranged and split into training, validation and test sets.

- Identify class (category) subsets with enough labeled images for training (at least 200)
- Balance number of images per class (for simplicity reasons) by random choice
- Exclude multi-labeled images for training set (for simplicity reasons)
- Extract the actual image size information: done once and stored in the RDF file linkable with the image metadata by the noticeRef.
- Exclude the images with the aspect ratio over 2.0 – too wide or too tall
- Split the image subset on train/validation/test (80%-10%-10%)
- Rearrange the image files to the structure suitable for the classifier's consumption.
- Rename the image files from the original to the associated unique noticeRef to simplify the post processing.

RUN THE CNN CLASSIFIER

[1] Some of the referenced images are not actually present in the provided image set

- As a baseline model use VGG16 with batch normalization [3] pre-trained on the ImageNet (PyTorch VGG16_bn model) dataset



- Transfer learning from the training of the network on the ImageNet dataset to decrease the training time. Learned parameters are available within the PyTorch framework.
- Image transformations
  - Resize to 256x256
  - Center Crop to 224X224
  - Normalization where the normalization values for means and standard deviation were calculated over the sample of images.
- Hyperparameters:
  - Mini batch = 4
  - Cross-Entropy loss function,
  - stochastic gradient decent (SGD) as optimizing algorithm
  - Learning Rate = 0.001, adjusted every 4 epochs
  - 10 epochs.
- Run training/validation in one of three modes:
  - only the last fully connected layer training
  - all fully connected layers training
  - all layers (full) training
- Select the final model based on the best validation accuracy of the iteration. Save the parameters in a file.
- Test the model on unseen images and record the metrics (precision, recall, f1 score per class and overall accuracy).
- Record the top-5 predicted probabilities of the classes for each test image and store as an RDF file.

POST-PROCESS OF THE CLASSIFICATION RESULTS

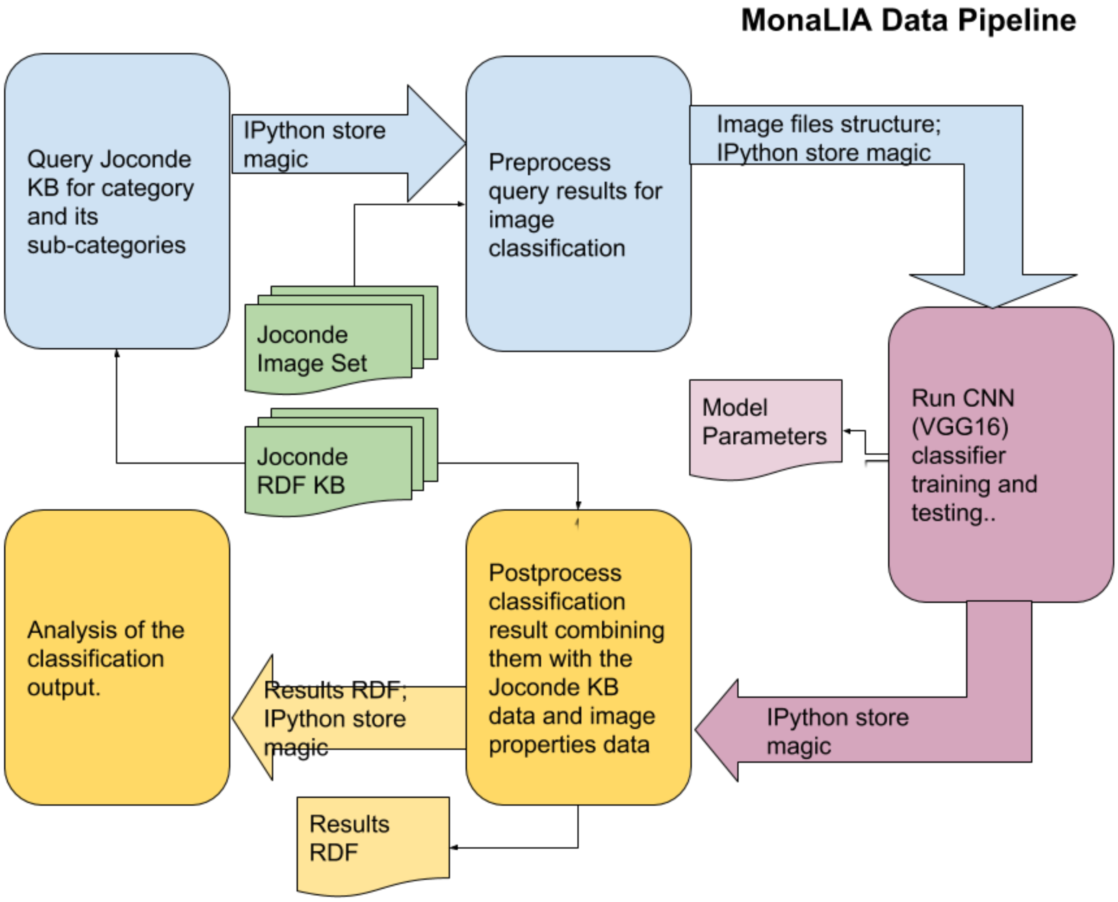
Query the KB metadata again to find the dependencies between the classification outcome and the artwork properties. Query the metadata only for the test images to reduce the data size.

- Link the acquired image size data with the metadata.
- Link the binary classification outcome (correct/incorrect) with the metadata.
- Exclude sparsely populated metadata variables (< 33%)
- Reduce the number of classes in the categorical variables (manually selected count thresholds for each variable)
- Run statistical analysis of the combined dataset to determine if there is any dependency of the prediction outcome and metadata variables
    - Cluster map of the confusion matrix to visualize the perceived similarities between the classes
    - Logistic regressions for the continuous variables (image width, height, aspect ratio)
    - Logistic regressions for categorical variables (art form, domain, technique, etc.)
    - Recursive Feature Elimination (down to 6)
    - Decision tree for selected variables to see if the outcome can be explained by the independent metadata and for the better visualization

IMPLEMENTATION

The steps described in the previous section are implemented as a set of 5 Jupyter Notebooks and a shared library script:

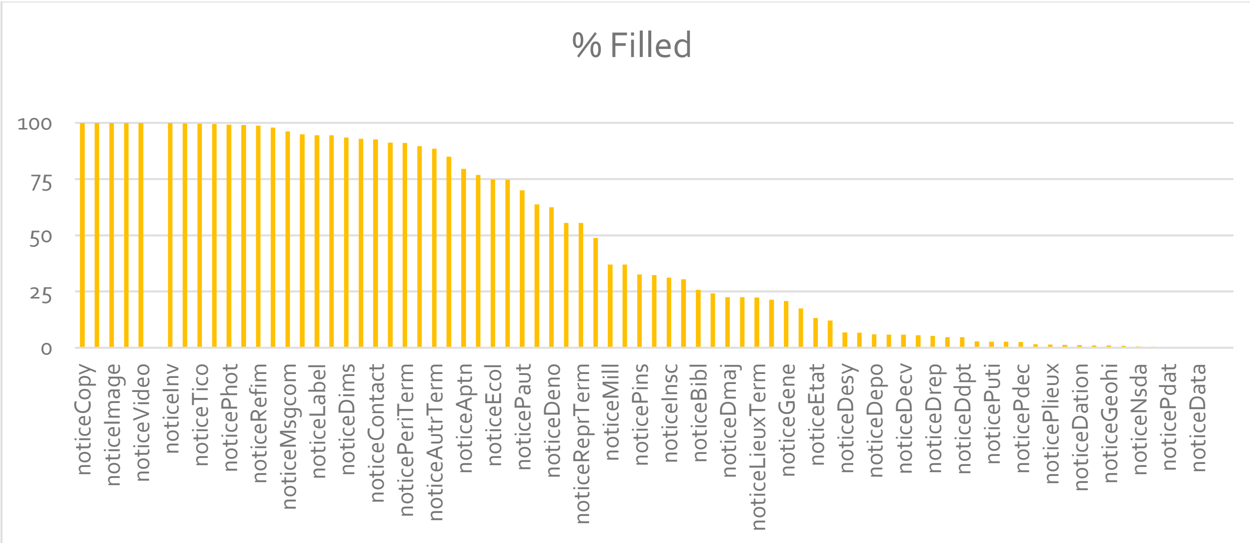
1. MonaLIA.STEP 1.KB Query Result To Dataframe.ipynb
2. MonaLIA.STEP 2.Image Set Preprocessing.ipynb
3. MonaLIA.STEP 3.Pretraned VGG16_bn Classification.ipynb
4. MonaLIA.STEP 4.Image Set Postprocessing.ipynb
5. MonaLIA.STEP 5.Classification Analysis.ipynb
6. MonaLIA.py



RESULTS

DATA QUALITY ASSESMENT

- 59% (285144) of all records (483297) have the images.
- Joconda ontology defines 76 properties of the artwork
  - 37% of the properties (28) are filled over 75%
  - 46% (35) are filled under 25%.
- 56% (165800) of the images have the represented subject associated with them.



- Some of the properties that might be useful from the data inference point of view have more than 1 entity in them with entities being entered in free form. That makes them hard to use in the automated processing. Below are the properties that may benefit from the indexing:

noticeLoca	Artwork location (city, museum)	99.9%	Contains both city and museum. However, the noticeMuseo property can be used instead.
noticePhot*	Artwork photo credit	99.2%	In many cases this field represent a company and a photographer
noticeLoca2	Artwork location 2	97,95%	Contains 2 or 3 entities representing the (country, region, department)
noticeTech*	Artwork technique/materials	94,55%	Can contain many entities inconsistently separated.
noticeDims	Artwork dimensions	93,44 %	Totally free texted. However this data might be very useful in image classification.
noticeDeno*	Artwork denomination	62,44%	Can contain many entities inconsistently separated

* Fields that were selected as explanatory variables to the image classification results analysis.

- A special attention was paid to the property *noticeReprTerm* that links the terms of the REPR thesaurus inspired by the iconographic thesaurus by Francios Garnier http://www2.culture.gouv.fr/documentation/joconde/fr/partenaires/AIDEMUSEES/thesaurus-garnier/thesaurus-pres.htm published in 1984.

All REPR terms	32274	100%

Terms associated with artwork records	29185	90%
Terms associated with artwork records that have images	22552	70%
Terms excluding names*	12013	37%
Terms associated with more than 200 images	790	2.4%

* Some 17426 terms are found that are grouped under the "quidams"@fr category. These terms are personal names and only assigned once or twice. These terms can be ignored.

The 10 most frequent terms are:

REPR Term	**Frfequency**	**% of images**
homme@fr	31709	10,62%
femme@fr	27754	9,29%
figure@fr	25323	8,48%
scène@fr	23755	7,96%
paysage@fr	23499	7,87%
portrait@fr	20859	6,99%
vue d'architecture@fr	14037	4,70%
ornementation@fr	11457	3,84%
en pied@fr	11099	3,72%
en buste@fr	10914	3,66%

- The terms are organized in the hierarchies with 12 top concepts. The levels of the hierarchies may not have the same semantic level across all or the sub-hierarchies. The deepest branches are 11 edges removed from the root.

Conclusion:

There is a room for improvement of the Joconde KB in terms of filling the missing data either by experts or by inference or by deep learning algorithms.

Some of the properties would benefit from splitting into several and indexing to make them more searchable and machine understandable. (for ex.: *noticeDeno*, *noticeTech*, *noticeEtat*)

Property *noticeDims* (artwork dimensions) is in free form and unusable. The dataset would benefit from formalizing the dimensions properties (width, height, depth) and the units (cm, mm, m, inch, etc.).

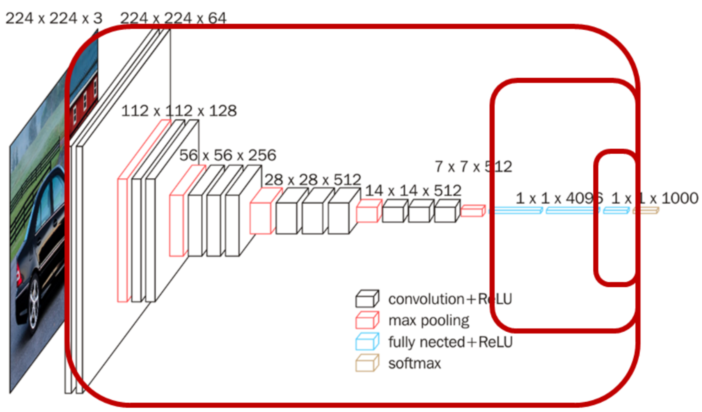
The number of the subject term is too large to be able to do the deep learning classification by one CNN.

By far not all of the subject terms have enough images to train the CNN to recognize the subjects.

TRANSFER LEARNING

Learned parameters obtained from the PyTorch framework for the VGG16_bn model trained on ImageNet and loaded on the model creation. Ran classifications in three modes:

1. Last fully connected layer only training
2. Fully connected layers only training
3. Full training (using the learned parameters as initial parameters)

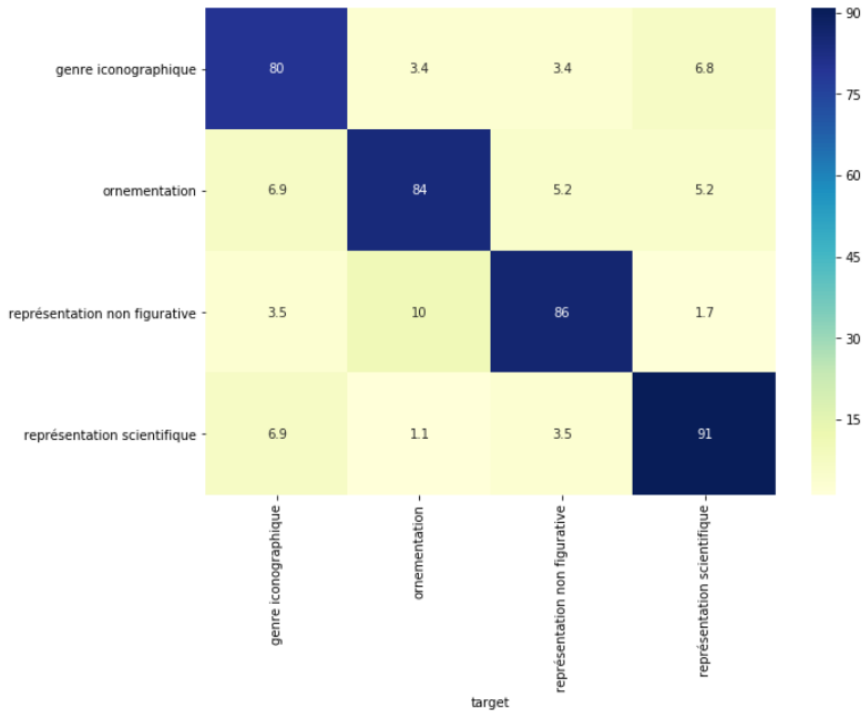


Performance details of the experiments on different image subsets:

REPRESENTATION TYPE (THEME)

- *genre iconographique*
- *ornementation*
- *représentation non figurative*
- *représentation scientifique*

Classification: Representation Type	**Number of classes**	**Train/Val/Test Set Size**	**Test Accuracy**	**Training Time (10 epochs)**
Last fully connected layer only training	4	1828 228 232	81%	10min
Fully connected layers only training	4	1828 228 232	81%	11min
Full training	4	1828 228 232	**85%**	54min



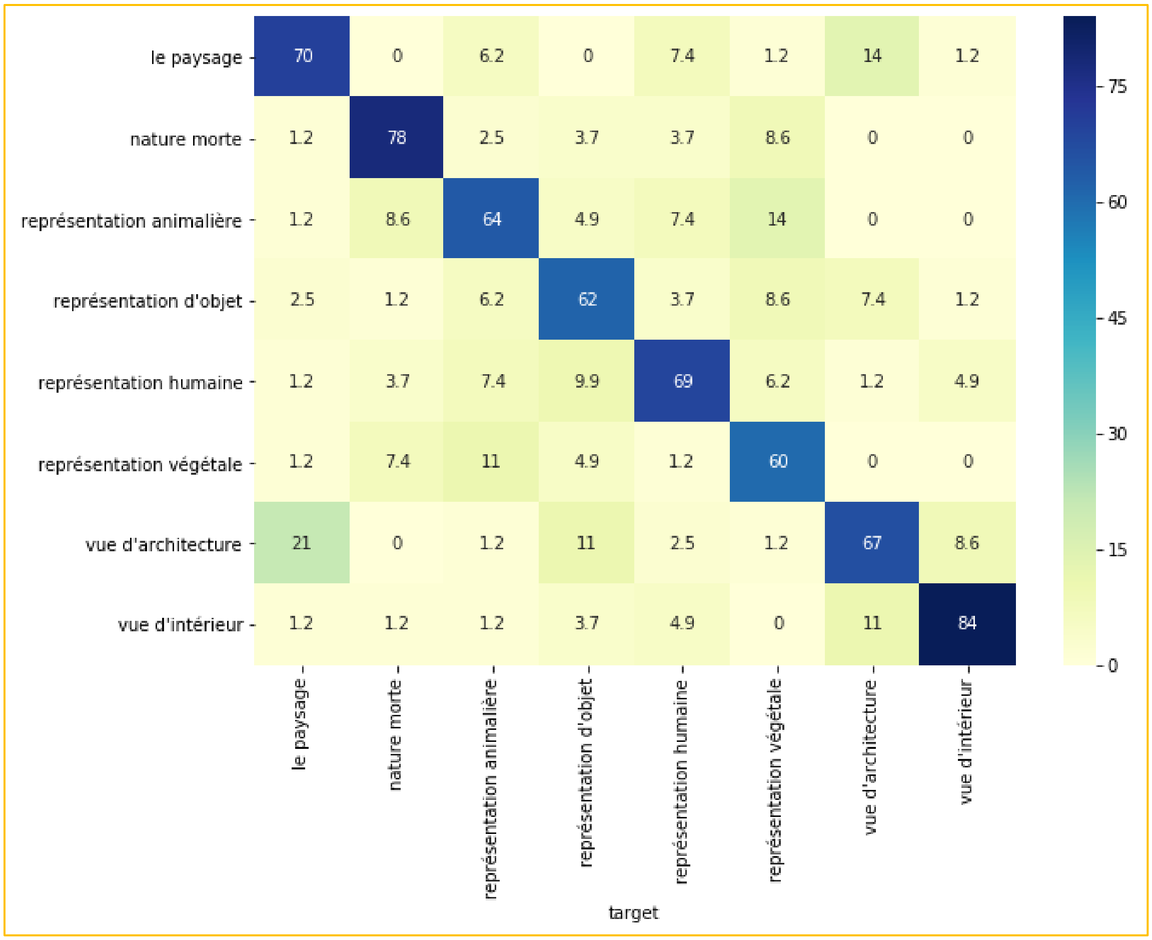
Confusion Matrix 1

ICONOGRAPHIC TYPES (THEME/REPRESENTATION TYPES)

- *le paysage*
- *nature morte*
- *représentation animalière*
- *représentation d'objet*
- *représentation humaine*
- *représentation végétale*
- *vue d'architecture*
- *vue d'intérieur*

Classification: iconographic types	**Number of classes**	**Train/Val/Test Set Size**	**Test Accuracy**	**Training Time (10 epochs)**
Last fully connected layer only training	8	5152 640 648	67%	30min
Fully connected layers only training	8	5152 640 648	66%	36min
Full training	8	5152 640 648	**69%**	2h 35min

Confusion matrix :

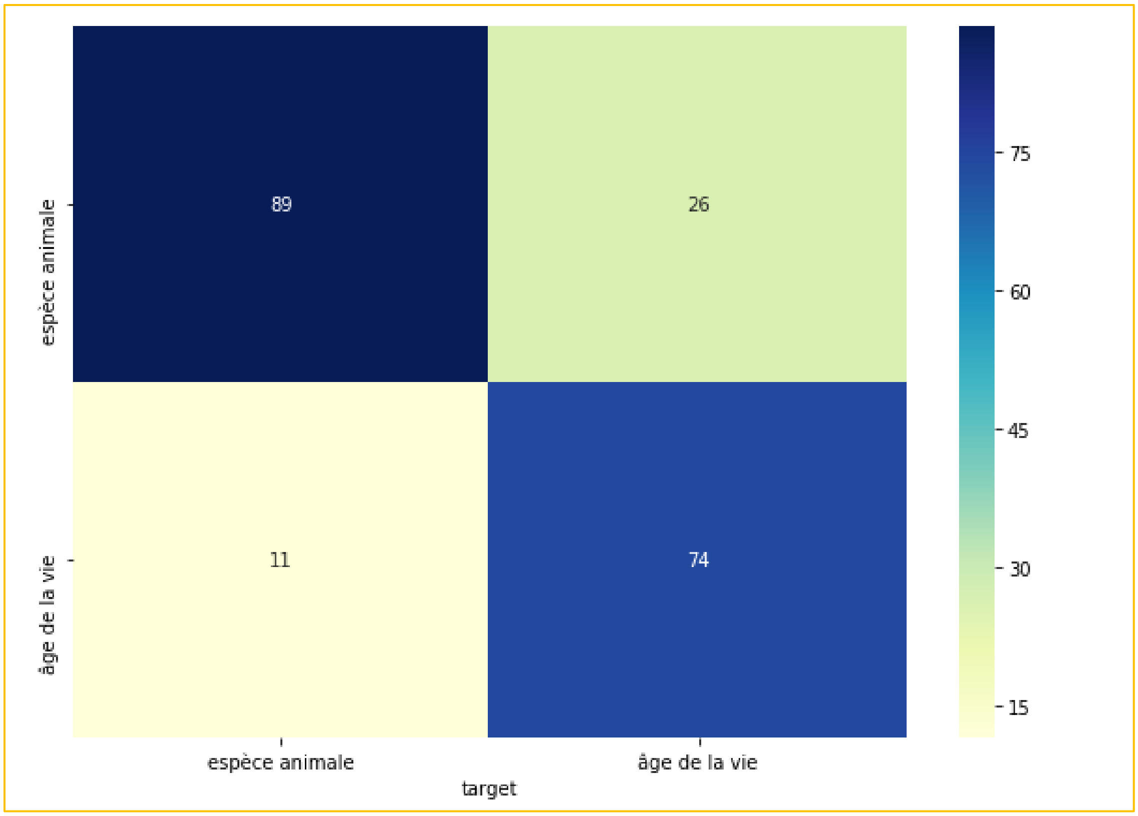


Confusion Matrix 2

ANIMALS VS. HUMANS

- *espèce animale*
- *âge de la vie*

Classification: animals & humans	**Number of classes**	**Train/Val/Test Set Size**	**Test Accuracy**	**Training Time (10 epochs)**
Last fully connected layer only training	2	23058 2882 2882	82%	2h 19min
Fully connected layers only training	2	23058 2882 2882	**85%**	3h 15min
Full training	2	23058 2882 2882		Not done

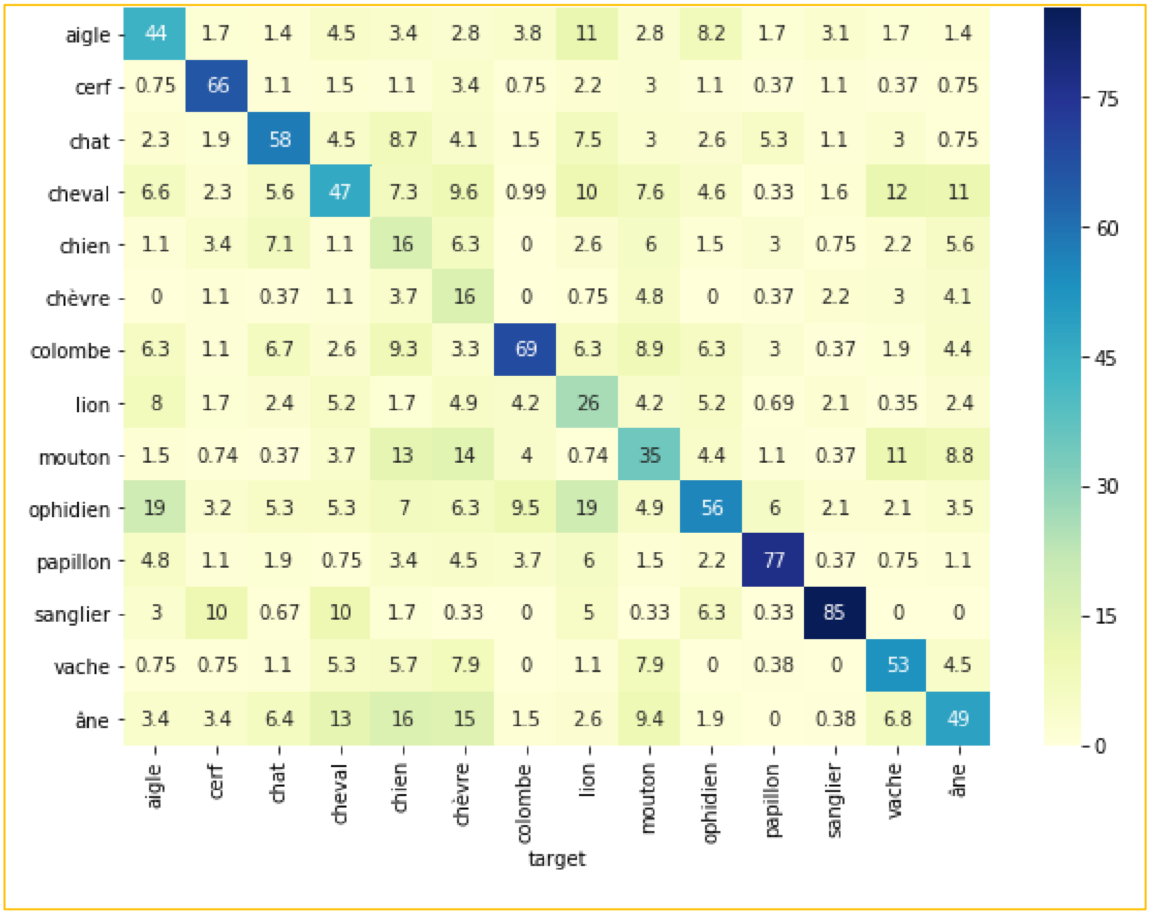


Confusion Matrix 3

ANIMALS

- *Aigle*
- *Cerf*
- *Chat*
- *Cheval*
- *Chien*
- *Chèvre*
- *Colombe*
- *Lion*
- *Mouton*
- *Ophidien*
- *Papillon*
- *Sanglier*
- *Vache*
- *âne*

Classification: animals	**Number of classes**	**Train/Val/Test Set Size**	**Test Accuracy**	**Training Time (10 epochs)**
Last fully connected layer only training	14	7153 893 893	32%	35min
Fully connected layers only training	14	7153 893 893	**40%**	1h 2min
Full training	14	7153 893 893	35%	3h 32min

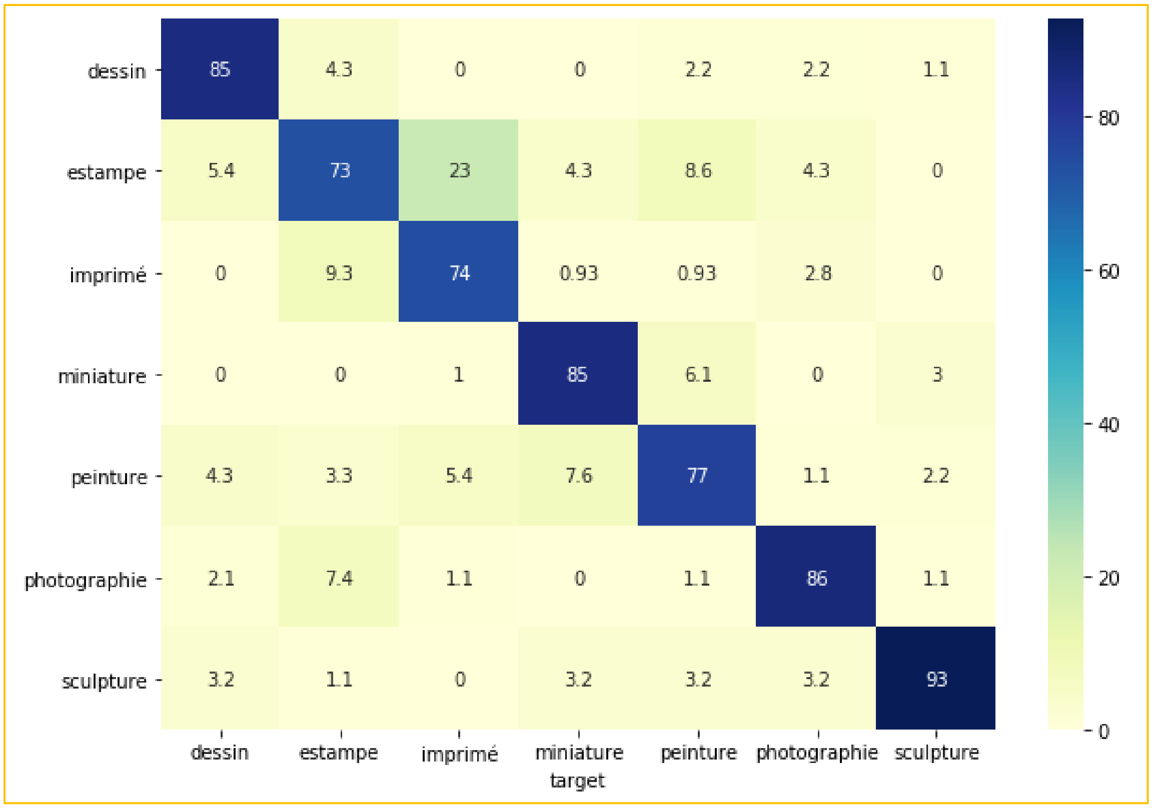


Confusion Matrix 4

ARTISTIC DOMAINS

- *dessin*
- *estampe*
- *Imprimé*
- *Miniature*
- *Peinture*
- *Photographie*
- *sculpture*

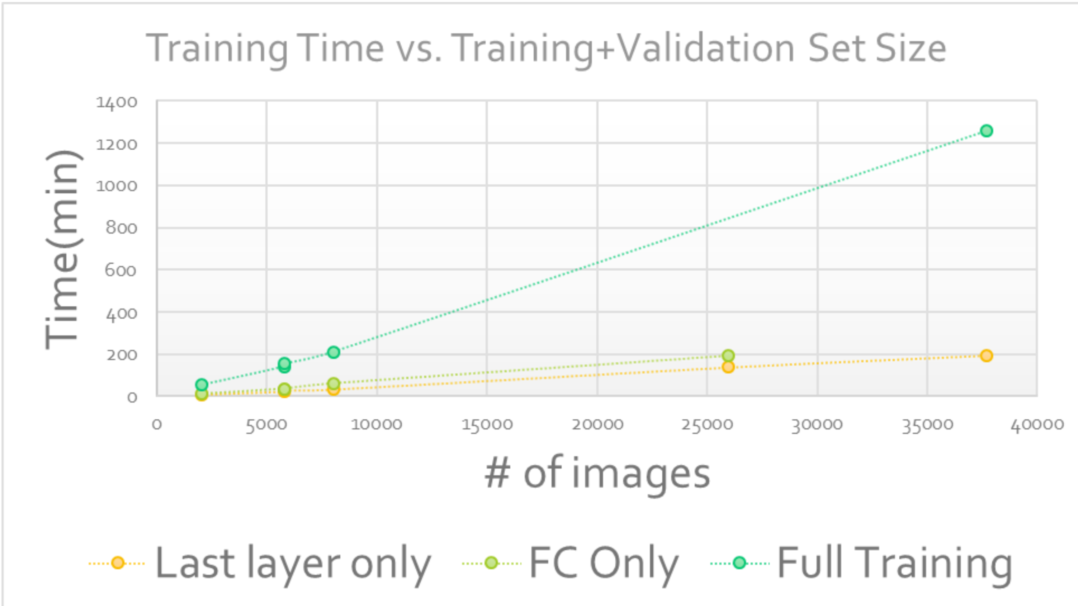
Classification: domains	Number of classes	Train/Val/Test Set Size	Test Accuracy	Training Time (10 epochs)
Last fully connected layer only training	7	5159 644 644	77%	25min
Fully connected layers only training	7	5159 644 644	76%	36min
Full training	7	5159 644 644	**83%**	2hr 22min



Confusion Matrix 5

Time performance summary on:

- Hardware:
  - Intel(R) i7 CPU @ 3.10GHz
  - 64 GB RAM
  - NVIDIA Quadro M2200 GPU
- Software:
  - Windows 10 64-bit
  - PyTorch v. 0.4.0
  - CUDA 9.1



Conclusions:

Transfer learning works pretty well saving time on model development and model training.

The transferred parameters are very good set of initial parameters so the initial learning rate can be chosen at 0.001 for the loss function and convergence reached within 10 epochs.

Full training improved the accuracy 2-3%, which means that the last layer only training gives reasonably good results in the fraction of the time.

The broader (semantically) the class the higher the accuracy due to the number of factors such as larger number of available training images, smaller number of classes. However, what is very interesting that the broad classes like iconographic types and domains do not correspond to the classes of the ImageNet and yet yield a good accuracy.

Another very interesting observation is that the network was able to distinguish between the photography as an art form from photographs of the other art forms.

For the Animals classification, the dense layer training performed the best in terms of both accuracy and time. My explanation is that the ImageNet dataset contains the same classes, so the features are the same and did not need to be retrained. In fact, it might be that the retraining based on the "worse" quality images makes the model less accurate.

IMAGE PRE-PROCESSING

At the beginning used the same image transformations that the VGG16 on ImageNet training used, that is RandomCropResize+ RandomHorizontalFlip+ Normalization. But the experiments shown that for the Resize(256) + CenterCrop(224) + Normalization the accuracy increased by 3%. Used the latter for the majority on the experiments.

- Calculated normalization vectors for 2 samples (animals, domains).
- Post-classification analysis shown that the image aspect ratio over 2 or under 0.5 has a negative effect on the classification accuracy.

Conclusions:

The experiments shown that there is not much difference in the classification accuracy between using the normalization values of the samples or even ImageNet normalization parameters.

The aspect ratio of the image file affects the image classification if the image dimensions are too far from square. Did not use the "long" or "tall" images for training. Filtered out the images with the aspect ratio greater than 2.

IMAGE SET AUGMENTATION

In the Joconde image set the distribution of the images of certain objects is not uniform thus the classes are unbalanced. To elevate the problem I downloaded the Painter by Numbers dataset from Kaggle that has 79434 images of artwork. If a certain class did not have enough training images, I've selected the images which had the class name in the title. For example adding images of the cats to the Cat-Dog-Horse classifier increased the test accuracy by 11%.

Classification: cats, dogs, horses	**Number of classes**	**Training Set Size**	**Test Accuracy** (last fully connected layer)	**Test Accuracy** (full training)
Training images only from the Joconde set	3	636	53%	57%
Training images augmented by images from Painter By Numbers	3	1662	**64%**	**68%**

Conclusion:

Adding the images from the external datasets for training increases the model quality.

CLASSIFICATION RESULTS ANALYSIS

An attempt of the statistical analysis was made to find the possible explanation of the miss-classifications. <u>The results are only preliminary and not exhaustive or conclusive</u>.

We were interested in what are the metadata properties might have an effect on the correctness of the image classification.

To answer this question I've ran logistic regressions and decision tree algorithms on the dependent binary variable (correct/not correct classification) and the preselected metadata properties. The properties were selected based on subjective relevance and sparsity. Some of the properties were split into multiple variables.

Binary	**Numeric**	**Categorical**
output (Label = Prediction)	imageWidth,	noticeArtForm (noticeDomainTerm)*
	imageHeight	noticeFunction (noticeDomainTerm)*
	imageAspectRatio	noticeDiscipline (noticeDomainTerm)*
		noticeRepresentationType (noticeReprTerm)*
		noticePhotocredit (noticePhot)*
		noticeMuseum (noticeMuse)*
		noticeTechnique1 (noticeTech)*
		noticeTechnique2 (noticeTech)*
		noticeTechnique3 (noticeTech)*
		noticeDenomination (noticeDeno)*

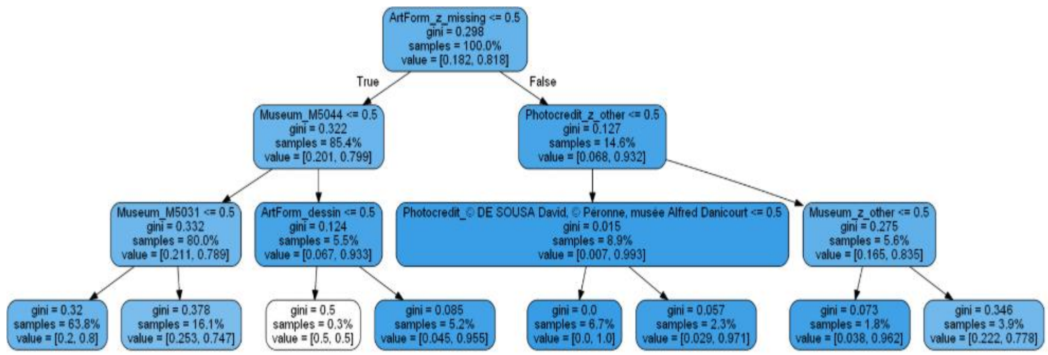
* Variable was created from the actual metadata property in parentheses.

For the continuous numeric variables, it appears that the size matters. Especially the aspect ratio of the image. With the increase of aspect ratio by 1 the positive outcome is reduced by 20%.

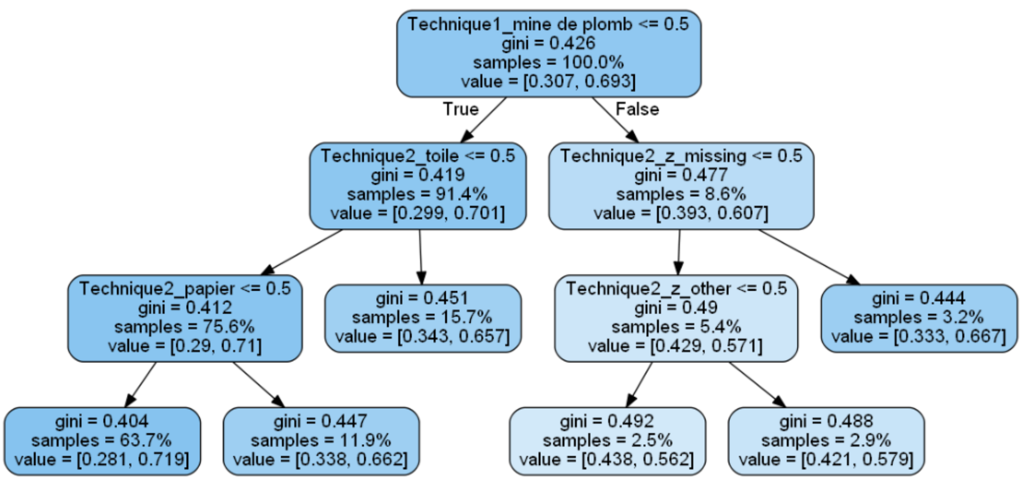
```
==============================================================================
ar
==============================================================================
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 4324
Model:                          Logit   Df Residuals:                     4323
Method:                           MLE   Df Model:                            0
Date:                Thu, 13 Sep 2018   Pseudo R-squ.:               -0.002428
Time:                        15:49:02   Log-Likelihood:                -2960.0
converged:                       True   LL-Null:                       -2952.9
                                        LLR p-value:                       nan
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.2336      0.027     -8.572      0.000      -0.287      -0.180
==============================================================================
exp(coef) = 0.7917
==============================================================================
```

For the categorical variables, the decision tree analysis shown that the classification outcome may be explained:

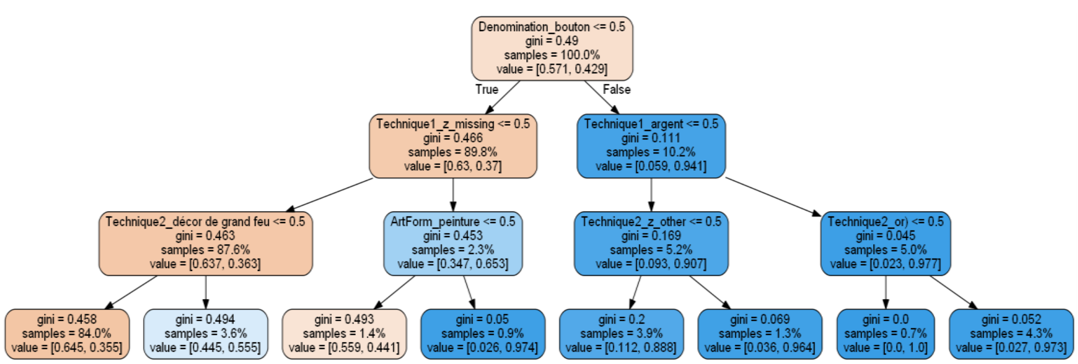
- by the art form, museum and the photography agency that took a digital picture. For example, these variables appear the most relevant for the Animals vs. Humans classifier:



- by the material of the artwork (paper, canvas, gold, etc.) . For example, this variable appear the most relevant for the animals and iconographic genre classifier:



- by the denomination (painting, button, etc.) in addition to the material (paper, canvas, gold, etc.) . For example, these variables appear the most relevant for the Animals classifier:



Conclusions:

It seems that there is a dependence of the classification outcome on some of the metadata properties. Unfortunately, the metadata properties identified are dirty and the concrete conclusion cannot be reached.

FURTHER RESEARCH

SEMANTIC INDEXING: JOCONDE KB

The most important decision has to be made on the subset of classes of interest out of 37279 from the REPR Thesaurus.

- Investigate more the hierarchies in the Garnier Thesaurus. Compare it hierarchy of ImageNet (based on WordNet).
- Investigate the other art description ontologies and feasibility of mapping them with Joconde ontology for further enhancement of the metadata. It may not contribute to the image annotation though.

  http://mappings.dbpedia.org/index.php/OntologyClass:Artwork

  https://www.loc.gov/standards/vracore/

IMAGE SET

- Depending on the eventual set of classes, that the project will be focused on we will need to create a "golden" image set for training. The accuracy of the training set should be high but without overfitting.
- Depending on the number eventual number of classes the dataset might need some augmentation:
  - Below are few other datasets that can be used to augment the training set. More external datasets can be explored.
    https://www.wga.hu/search.html
    https://www.wikiart.org/ with https://github.com/lucasdavid/wikiart
    https://www.nga.gov/collection/collection-search.html
    https://bam-dataset.org//

  - Research if applying the artistic transformations to the ImageNet images can be the way to augment data. This might be e good starting point:
    https://github.com/leongatys/PytorchNeuralStyleTransfer,
    https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Gatys_Image_Style_Transfer_CVPR_2016_paper.pdf

DEEP LEARNING: IMAGE CLASSIFICATION

- Investigate the different image transformation techniques available in PyTorch torchvision package https://pytorch.org/docs/stable/torchvision/transforms.html to improve the classification accuracy.

- Build the multi-label classification model to detect multiple objects on the image. The majority of the artworks represent more than one subject therefore it will be interesting to investigate the possibility of detecting multiple classes on the same image using the multi label training techniques like One-hot-encoding. https://towardsdatascience.com/multi-label-classification-and-class-activation-map-on-fashion-mnist-1454f09f5925

- Build a hierarchical classification model. Investigate the depth of the hierarchy that can be useful in image annotation. Investigate the possibility of using the hierarchy in recognition of the different levels of details. For example, if the species of the animal cannot be recognized with high enough probability (cats, dogs, etc.) but the family is more certain (Felidae, Canidae, etc.) then the description of representation can be chosen just on the family level.

- Build a model to recognize the theme and/or iconographic genre and the content of the image by the same CNN. Investigate if the theme recognition can help with the object recognition. For example, train the classifiers for the iconographic genre (human representation, animal representation, etc.) and more specific subjects (man, woman, animal, bird, etc.) would be weighted accordingly.
- Investigate the possibility of using the RNN classification layers based on this paper https://link.springer.com/article/10.1007/s11042-017-5443-x.
- In the preliminary studies, we only used VGG16 model. Investigate if the other well-known pre-trained models can improve the prediction accuracy. https://pytorch.org/docs/stable/torchvision/models.html
- For the promising models build in the cross validation to determine the prediction power of a model.

APPENDIX A: PYTHON PACKAGES

Creating and querying the KB graphs

- Rdflib
- SPARQLWrapper

Deep Learning

- torch
- torchvision

Statistical analysis

- sklearn
- statsmodels

Data visualization

- matplotlib
- livelossplot
- pydotplus, IPython
- seaborn

Data manipulation

- numpy
- pandas

APPENDIX B: ADDITIONAL RESULTS FILES

MonaLIA.Classification Samples.xlms.	The sample results of the classification
MonaLIA.VGG16-BN Performance.xlsx	The metrics of all the CNN runs.
MonaLIA.missclassifications research.xlsx	Joconde KB properties analysis
MonaLIA.REPR Research.xlsx	Statistics on REPR thesaurus of Joconde KB
MonaLIA.DOMN Research.xlsx	Statistics on DOMN thesaurus of Joconde KB

REFERENCES

[1] Liu, Q., Jiang, H., Evdokimov, A., Ling, Z. H., Zhu, X., Wei, S., & Hu, Y. (2016). Probabilistic Reasoning via Deep Learning: Neural Association Models. arXiv preprint arXiv:1603.07704.

[2] Johnson, J., Ballan, L., & Fei-Fei, L. (2015). Love thy neighbors: Image annotation by exploiting image metadata. In Proceedings of the IEEE international conference on computer vision (pp. 4624-4632).

[3] K. Simonyan, A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition.Proceedings of ICLR 2015, p.1-14.

[4] Kuang, Z., Li, Z., Zhao, T., & Fan, J. (2017). Deep Multi-task Learning for Large-Scale Image Classification. 2017 IEEE Third International Conference on Multimedia Big Data (BigMM), 310-317.

[5] Kuang, Zhenzhong & Yu, Jun & Li, Zongmin & Zhang, Baopeng & Fan, Jianping. (2018). Integrating Multi-Level Deep Learning and Concept Ontology for Large-Scale Visual Recognition. Pattern Recognition. 78. 10.1016/j.patcog.2018.01.027.